Chapter 10
Ontology Learning and the Humanities

Teddy Bunn
The University of Western Australia, Australia

ABSTRACT
This chapter reviews the current state of play in the use of ontologies in the humanities, with best-practice examples from selected disciplines. It looks at the specific domain problems faced by the humanities, and examines the various approaches currently being employed to construct, maintain, and develop humanities ontologies. The application of ontology learning in the humanities is discussed by reviewing a range of research projects in different disciplines. The chapter concludes with an assessment of the future potential of ontology learning in the humanities, and an attempt to set out a research agenda for this field.

INTRODUCTION
The humanities are academic disciplines which study the nature of human life and experience. They are different from the natural and social sciences because they use methods which are mainly analytical, critical, or speculative. There are various different definitions of the scope of the humanities. According to the Australian Academy of the Humanities they cover the following disciplines: Archaeology; Asian Studies; Classical Studies; English; European Languages and Cultures; History; Linguistics; Philosophy; Religion and the History of Ideas; Cultural and Communication Studies; the Arts.

Ontological frameworks are central to the work of humanities researchers. This is because most humanities research involves either the analysis and definition of concepts or the categorization of individual phenomena into broader classes. Philosophy is the pre-eminent academic discipline which focuses on concepts, while the focus of the disciplines of history and archaeology is largely on the categorization of specific instances (people, places, events, objects and so on). Other humanities disciplines rely on a mixture of these two approaches.

This paper reviews the current state of play in the use of ontologies in the humanities, with best-practice examples from selected disciplines. It looks at the specific domain problems faced by the humanities, and examines the various approaches currently being employed to construct, maintain and develop ontologies.

The applications of ontology learning in the humanities is also examined, by reviewing a range of research projects in different disciplines. Areas discussed include the availability of text corpora and other sources of knowledge, and the use of text mining techniques and tools. The standards and tools used for expressing and developing ontologies are also covered.

The paper concludes with an assessment of the future potential of ontology learning in the humanities, and an attempt to set out a research agenda for this field. It also aims to identify areas where ontology learning is likely to prove most valuable and applicable.

Ontologies and the Humanities
The humanities are particularly difficult area for the development and application of ontologies. The semantic context is complex and often ambiguous, and there are numerous existing vocabularies and taxonomies which often overlap. The multilingual nature of much humanities research has obvious implications for the use of ontologies. Not only are there different terms in different languages for the same concept; a term may be similar in appearance across different languages but have quite different meanings. Humanities research also often crosses different time periods, and the meanings of words and concepts tend to shift over time. Other worldviews and ways of classifying the world – both current and past – also need to be factored in. Indigenous knowledge (the Australian aboriginals or the North American Indians, for example) has a radically different hierarchy of classes, in addition to the obvious linguistic differences. A further area of complexity is the enormous proliferation of instances, particularly of people and places.

Nagpal (2005, pp. 208-209) discusses four specific difficulties:

- **Time dependence:** almost every instance is time-dependent, e.g. the content of a book is different from one moment to the next. This means that the relationship between an instance and its class is likely to be different at different time periods.

- **Uncertainty:** the documentation relating to instances is frequently missing or contradictory, especially about dates. There are at least two different recorded birth dates for Joseph Stalin, for instance: 21 December 1879 in the official Soviet records, and 6 December 1878 in the church records. In another typical example, the paintings of the Dutch artist Vermeer cannot be dated precisely. This kind of uncertainty may affect the relationship between an instance and its class.

- **Subjectivity:** “most complex historical notions are vaguely defined or open to multiple interpretations, and thus can be interpreted subjectively, making them difficult to model conceptually.” Concepts like the Enlightenment or the Middle Ages do not have precise beginning and ending dates, for example, and it can be unclear whether a given temporal instance falls within the scope of a concept like this. Opinions and interpretations can differ considerably, even between experts.

- **Why?** questions: historians tend to focus on this kind of question, and are interested in seeing facts in an explanatory context.
Ontology Learning and the Humanities

rather than on their own. Ontologies are much better at representing precise facts, and historical knowledge is particularly difficult to model in this way.

These difficulties have led some researchers to question the very applicability of ontologies in the humanities (Velmans 2004). Nevertheless, there has been a growing body of research in recent years which is aimed at the development and maintenance of such ontologies. There are also an increasing number of digital resources for the humanities which make use of ontologies, especially in the field of knowledge discovery for cultural heritage collections. Mark Greenmeyer (2007) has identified three basic approaches to the development of humanities ontologies:

- Top-down ontologies: generated from taxonomies accepted within a discipline;
- Middle-out ontologies: generated by intelligent iteration;
- Bottom-up ontologies: generated from a representative sample of canonical data.

Most of the work in the humanities to date has concentrated on the production of top-down and middle-out ontologies.

Two important examples of broad top-down ontologies relevant specifically to the humanities are VICO and CIDOCCRM. VICO (Visual Contextualization of Digital Content) was a European Union project which developed an ontology to serve as the basis for searching a European history portal. The VICO ontology was developed using a middle-out iterative approach. A bottom-up approach was not thought to be possible because the project could not identify a suitable textual corpus to use for deriving such an ontology. The VICO ontology consists of only seven basic concepts (or "flavours"), with a shallow hierarchy of two or three levels. This structure is designed to allow the population of the ontology with large numbers of instances and relationships. Over 15,000 instances were uploaded from Excel spreadsheets, though this work took about two years to carry out. The VICO ontology instances referring to historical people and events have subsequently been translated into twelve European languages as part of the ENRICH project, which has been implementing a Web catalogue of medieval manuscripts.

The VICODI project encountered two major difficulties (Gravegna, et al., 2008, p. 71). The first was that there is no accepted corpus of documents or texts from which a conceptual framework could be developed. There is also no organization with sufficient authority in the field of historical research to promulgate or sponsor the use of a particular scheme of knowledge classification. The second major difficulty was that there are major differences between the terminology used to write about history and the terminology used in the historical sources themselves. The project team concluded that "the VICODI ontology development process has shown that a complex humanities domain can be represented through a shallow ontology structure and a limited number of concepts and properties." The project left open the question of whether "constructing an ontology for any humanities domain is too labour intensive and too costly" (Nagyirp, DeWarte & Oosthoek, 2005, p. 346). It also led to the conclusion that "it is practically impossible to build a monolithic ontology for such a complex domain as history." (Nagyirp 2005, p. 213)

The CIDOC Conceptual Reference Model (ISO 21127:2006) is a core high-level ontology for describing concepts and relationships used in cultural heritage documentation (Doerr, 2003; Gill, 2004). It was developed as a top-down ontology by Martin Doerr and his colleagues at the Foundation for Research and Technology (FORTH) in Heraklion, Greece. In its latest version, it defines 90 classes and 148 properties. It was designed to enable semantic interoperability between the disparate sources of information produced by libraries, archives, museums and galleries. CIDOC CRM has been used as "semantic glue" by a range of services and projects, mainly in the field of museum documentation. Among these is CLAROS (Classical Art Research Online Services), which combines more than two million database records for Greek and Roman art objects held in a variety of museums.

Because of its universal and high-level approach, the CIDOC-CRM is also being mapped to knowledge structures in closely related domains. The Text Encoding Initiative (TEI) is widely used as the accepted XML schema for text encoding in the humanities. A draft mapping of TEI elements to CIDOC-CRM was completed in 2001, and there is considerable current interest in modelling vocabularies within specific TEI elements using the CIDOC-CRM framework (Oomens & Eyle, 2009). Another mapping is with FFBR (Functional Requirements for Bibliographic Records) - the internationally accepted XML schema for encoding bibliographic descriptions, which is being used to re-engineer the systems used by libraries to catalogue published materials. FFBR has been harmonized with CIDOC-CRM and transformed into an object-oriented version, FFBRoo (Riva, Doerr & Zimmer, 2009).

In addition to high-level ontologies such as VICO and CIDOC CRM, a wide range of more specific ontologies have been developed and used by humanities projects and services. One important example is the Museum Inland project and its successor CultureSamp in the Netherlands, which built a sophisticated demonstrator system with more than 200,000 cultural heritage objects from six different cultural heritage collections (Schröder et al., 2008). None of these ontologies were developed for this demonstrator. Four existing vocabularies were re-used instead: WordNet, and the three thesauri maintained by the Getty Research Institute – the Art & Architecture Thesaurus (AAT), the Union List of Artist Names (ULAN) and the Thesaurus of Geographic Names (TGN). These vocabularies were made interoperable through their conversion to the SKOS format and subsequent alignment using owl:sameAs and skos:exactMatch relations. The demonstrator has been limited to “simple syntactic alignments” aimed at boosting search results.

Other humanities research projects have developed and applied ontologies on a smaller scale. Mirzaee, Iversen and Hambidge (2005) provide a useful example of an iterative middle-out approach to ontology-building, applied to a single book on the history of the Islamic constitution. They adopted the following steps:

1. Identifying the purpose, scope, and users
2. Domain analysis and knowledge acquisition
3. Building a conceptual (informal) ontology model
4. Formalization
5. Evaluation

Ontology Learning and the Humanities
They drew on various techniques: competency questions and scenarios, brainstorming, informal analysis of the text, other general ontologies and ontology libraries. After designing, building and formalizing their ontology, they populated it with more than 750 instances extracted from the history book. The same competency questions methodology was also used by the ONTO-H project to build a general humanities ontology for annotating texts, designed as a plugin to the Protégé software (Benjamin, et al., 2004).

Other humanities projects have reported using ontologies but have not described the methodology used to develop their ontology. The Lunda of the Normans project, for example, applied a small-scale ontology to build a database of persons, places, source documents and relationships relevant to the land-holdings of the Norman aristocracy in twelfth-century England (Power, et al., 2007). But the methods used to build this ontology are not described in the project documentation. The report on the research by Gijs Geleijne and Jan Korei (2007) aimed at populating an ontology on historical persons by using information extracted from the Web does not explain how the ontology itself was developed.

Ontology Learning and the Humanities

The complexity of the linguistic and conceptual environment for the humanities has made ontologies difficult research field. It is hardly surprising, therefore, that the application of ontology learning techniques in this domain has been slow and infrequent. Four recent projects have broken important new ground, however, and shown considerable promise for the future. These projects are based on major textual corpora in different humanities disciplines: classical literature, 18th-century French writings, 13th-century Latin documents, and 18th-century accounts of trials.

Perseus Digital Library

The Perseus Digital Library is a major collection of literature and related resources in classical Greek and Latin, developed and maintained by a research team which has played a leading role in applying new semantic technologies to humanities data (Crane, Seales & Terras, 2009). One of its recent projects has been to investigate new approaches to the representation of fragmentary texts in the digital environment (Romanello, Bertin, Bossi, Béroud, & Crane, 2009). These findings only survive as quotations or references embedded in the writings of later authors. They pose a particularly difficult challenge for scholars in identification, attribution, and dating, and hence in modelling and digital representation. This project involved designing an ontology to represent these fragmentary ancient texts and to make them available in RDF as linked data on the Web.

To design this ontology, an initial analysis of the knowledge domain was carried out using ontology learning techniques. A corpus of 170 research articles was assembled from journals on classical philology contained in the JSTOR archive. The selection of the articles was made by a philologist with particular expertise in the history of fragmentary texts. The articles were in English and covered a range of different genres. These other specialist scholars worked with the project to assess and filter the results of the ontology learning process.

Initial pre-processing of the text corpus was done with the TreeTagger software, which tagged parts of speech and produced lemmatizations. The texts were then processed with Infomap, which was used to apply Latent Semantic Analysis. “Fragment” was chosen as the first seed term, and the Infomap associate tool was used to identify the terms most closely associated with it. These categories of terms were then classified to form the second generation of seed terms:

- philological topics (e.g. “quotation”);
- subjective evaluation and uncertainty (e.g. “supposition”); and,
- relationships – whole/part or spatial (e.g. “beginning” and “end”).

From this, new seed terms were generated iteratively until no relevant new terms were provided. The semantic relationships were then reduced to two dimensions and clustered through the use of the k-means algorithm, enabling their representation in a two-dimensional graph. The resulting ontology of concepts related to “fragment” was then linked into a broader ontological framework. A new class (“textual interpretation”) was created within the PhiloSURFsical ontology of philosophical concepts, as a sub-class of the “interpretation” class. Authors (conjectural and certain) were modelled as part of the FOAF ontology (as instances of foaf:person), and also as part of the CICOC-CRM ontology for cultural heritage objects (as instances of the “name” sub-class of the “appellation” class). The evidence for the fragment (i.e. the text and edition in which it was cited) was modelled through the BIBO ontology for bibliographical entities (as instances of the “book” class).

Using an ontology of this kind to represent fragments of classical texts was considered to be successful and valuable. The method used to build the ontology “demonstrated the importance of basing the ontology design on evidence that spontaneously emerge from a text corpus” (Romanello, et al., 2009, p.171).

Diderot’s Encyclopédie

The Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers was published in Paris between 1751 and 1772, under the direction of Denis Diderot and Jean le Rond d’Alembert. It originally appeared in a total of 28 printed volumes, and contains more than 7,000 articles covering all areas of human knowledge. The articles appear in alphabetical order, but most of them are also assigned to categories and sub-categories in a scheme for the classification of knowledge. About 70% of the articles are classified in this way.

A project carried out by staff of the ARTFL Project at the University of Chicago used the Encyclopédie as a test case with supervised learning algorithms (Horton, Morrissey, Olsen, Roe, & Voyer 2009). There were three initial objectives:

- To train a classifier on the classified articles in the Encyclopédie and then apply it to classify the articles which were left unclassified in the original publication;
- To reclassify the classifier to the classified articles and compare the results to the original classifications; and,
- To apply the classification model to other French literary texts of the eighteenth century and to evaluate its applicability beyond the Encyclopédie.

Pre-processing of the data involved extracting the text of all of the classified articles and removing the classification terms present at the beginning of these articles. TreeTagger was used to tokenize and lemmatize the texts, and frequencies for words and lemmas were computed. A Multinomial Naïve Bayesian classifier was then applied to the corpus. It was found to work most effectively when articles of 25 words or fewer were excluded, together with words which occurred in less than four articles. The classifier was then applied to the 22,000 unclassified articles from the original text. Manual analysis of a sample of results led to the assessment that the classifier had performed “reasonably well”.

In the final stage of this project, the classifier was used to re-classify all 54,289 originally classified articles. The classifier succeeded in assigning the “correct” classifications to 71.4% of the articles. This success rate could barely have been improved by some additional techniques, but it also reflects some anomalies in the original clas-
With the need to represent its historical context and interpretation" (Ciula, Spence & Vieira, 2008, p. 323). They were less convinced that the project had been able to develop an ontology which could be reused by other projects. The specific structure of the Fine Rolls was likely to limit its applicability beyond similar historical sources of the same general period.

Armadillo and Distributed Historical Sources

Armadillo is a software agent designed by the Natural Language Processing Group at the University of Sheffield. Its purpose is to provide machine readable content from large repositories for the Semantic Web, using automated methods. Doing this through manual or semi-automated annotation would be extremely costly and time-consuming. Techniques for information extraction are also unsuitable, because the documents have not already been annotated. Instead, Armadillo annotates texts automatically, with little or no manual intervention. It extracts information from a range of different sources and integrates this into a repository, from which it can be searched as well as linking back to or annotating the original source. These links ensure that the provenance and accuracy of the extracted information can be verified (Caragea, 2008).

Armamidillo relies primarily on redundancy — i.e., the fact that the sources contain multiple evidence for the same facts, in formats which are superficially different. Because a system knows the name of an author, for example, it can be used to identify other authors' names in digital resources, instead of relying on rule-based or statistical approaches, or manually constructed lists. By using this redundancy strategy, Armadillo can model the relevant domain, build an RDF ontology and a knowledge base, and connect findings across a corpus of distributed resources.

Armadillo was initially tested with technical documentation relating to jet engines and emergency response. A project carried out between 2005 and 2007 then aimed to test its applicability to 13th-century historical sources. Its History demonstrator draws on five digital documentary collections: the Old Bailey Proceedings Online, fire insurance policies of the Sun and Royal Exchange, wills registered in the Prerogative Court of Canterbury wills, the Settlement Examinations for the parish of St Martin-in-the-Fields, and the Westminster Historical Database. The most important of these is the Old Bailey Proceedings Online, which contains the published accounts of more than 100,000 trials in the Old Bailey court in London between 1674 and 1834.

The first four of these sources are relational databases, each with its own defined set of tables. The Old Bailey Proceedings, on the other hand, are an XML encoded text repository, with more than forty tags and attributes. The problem faced by the project was how to integrate these different approaches into a single ontology. The first step was to generate RDF from each of the datasets, using XLST scripts to retain the underlying structures in the data. Rather than mapping these to a single pre-existing ontology like VI-CODI, however, the Armadillo project mapped the various fields into which the data had been divided, to form a relatively simple ontology for each dataset. Each of these was then linked to enable cross-searching.

Because of time constraints, several fields in the datasets were not mapped into the ontology — notably those relating to time and time-intervals. This was one of the major weaknesses of the demonstrator identified during a subsequent independent evaluation carried out by graduate students and data developers. While the Armadillo history demonstrator was acknowledged as being a very limited first step, it produced sufficient evidence to support the general value of using an ontological framework for linking distributed historical datasets.
FUTURE RESEARCH DIRECTIONS

Within this broad field, some more specific questions can be identified as constituting a future research agenda. The relationships between ontologies and text encoding is an interesting area which models literary characters according to their mental representations. This is a semi-automatic process which reduces the need for manual extraction and classification of data. There are numerous textual corpora which are already in TEI encoding. Adding to their value through the use of ontology learning techniques will ensure that they remain a central component of the future Web of humanities data.

Zöllner-Weber’s project also reflects another major area of research interest: the use of automated methods to identify instances and populate existing ontologies with them. Geleijne and Korst (2007) describe a method of populating an ontology on historical people with information extracted from search engine query results. The first step on this process was to build a collection of snippets of text in response to Web queries. These snippets were analysed to identify names of people using a rule-based approach. An approach based on machine learning was not used, for three reasons:

- No representative training set was available. The corpus was fragmentary and multilingual.
- The corpus consisted of uncontrolled texts.
- The task of named entity recognition was simplified by the occurrence of patterns—primarily that a person’s name preceded a pattern of dates (birth—death).

The results obtained from this rule-based approach compared very favourably with results obtained by using the Stanford Named Entity Recognizer.

Matching between different ontologies is another important area for future research, given the proliferation of vocabularies in the humanities and the work being done to convert them to ontology-like formats. Improving techniques for ontology matching, especially across different languages and different disciplines, has been the focus of several recent projects. Some work has already been done on using semi-automatic methods of identifying links between vocabularies, using art history texts (de Boer, van Someren & Withgots, 2002). Two related Dutch projects also tested ontology matching techniques against cultural heritage vocabularies used by the Koninklijke Bibliotheek in the Netherlands and the Bibliothèque nationale de France (Isaac, et al., 2009; Angeli, et al., 2009).

Designing ontologies to cope with concept shifting over time is another area of great relevance to the humanities. Museum-Finland is one of the few projects to date which has successfully addressed this problem. A spatiotemporal ontology of Finnish counties between 1865 and 2005 was developed as one of Museum-Finland’s products. The data was designed to enable information retrieval which takes into account more than 600 changes in the borders and names of local administrative areas in Finland during this period (Kauppinen & Hyvönen, 2007).

More large-scale textual resources are becoming available, accompanied by new search and analysis tools. There will be an increasing opportunity and incentive to apply data mining techniques to extract ontologies from these kinds of corpora. The recent “Digging into Data Challenge”, sponsored jointly by the National Science Foundation (NSF), the National Endowment for the Humanities (NEH) and the Jisc Information Systems Committee (JISC), is a sign of things to come. Eight projects are being funded to analyse large datasets in a variety of formats: texts, images, audio recordings, music, geographical and numerical. They include teams working with the Ptolemaic Digital Library and the Old Bailey Proceedings (described above). The JISC-funded Connected Histories Project, being led by the University of Sheffield, is extending the work done by the Armadillo project (described above) to a total of fourteen structured and unstructured textual datasets.

A major potential impediment is the commercial ownership of important text corpora (like Early English Books Online) and existing vocabularies (like those produced by the Getty Research Institute). Geleijne and Korst (2007, p. 156) start with the assumption that “when ... all knowledge available on a domain can be found on the Web, this corpus can be used for ontology-driven information extraction.” This is not the situation in many humanities fields, where key datasets and corpora are owned by commercial firms and are not freely available for ontology-related research and experimentation. This is likely to hinder future experimentation with ontology learning in the humanities.

CONCLUSION

There is a growing recognition that ontologies are vital for building next-generation Semantic Web services aimed at the humanities research community. This is despite the well-documented difficulties involved in applying ontologies to this knowledge domain. Some work has been done on developing general humanities ontologies using top-down and middle-out methodologies. VICOI and the CIDOC CRM are important examples of this process. But most research to date has gone into converting existing vocabularies and thesauri into ontology-like formats, particularly RDF and SKOS, as well as into identifying ways of matching related vocabularies.

As a result, ontology learning is still in its infancy in the humanities. Several recent projects have experimented with the application of data mining techniques to textual corpora. For the most part, this work has emphasized using machine learning and data mining to identify instances to populate a limited range of classes, rather than building an entire ontology from scratch. The results have generally been quite encouraging. The ARTIFL researchers working on the Encyclopaedia project drew the conclusion that “traditional humanistic inquiry can be enhanced and broadened through the judicious application of machine learning and data mining techniques” (Horton, et al., 2009).

REFERENCES


Ontology Learning and the Humanities


**KEY TERMS AND DEFINITIONS**

Cultural Heritage: The artefacts and attributes of a society which are inherited from past generations.

Humanities: Academic disciplines which study the human condition.

Thesaurus: A type of controlled vocabulary which shows the hierarchical relationships between terms.

Controlled Vocabulary: A list of words and phrases for use in information retrieval, showing authorized and preferred terms.

Bibliographical: Relating to the description and study of books and other publications.

Text Encoding: The representation of texts in digital form using a markup language.

Text Corpora: A large and structured collection of texts for linguistic analysis.

**ENDNOTES**
